

Measurement-based Experimental Research Methodology

Dimitri Papadimitriou¹, Lluís Fàbrega², Pere Vilà², Davide Careglio³, Piet Demeester⁴

Alcatel-Lucent Bell Labs, Antwerp, Belgium
dimitri.papadimitriou@alcatel-lucent.com
Universitat de Girona, Girona, Spain
{lluis.fabrega,pere.vila}@udg.edu
Universitat Politècnica de Catalunya, Barcelona, Spain
careglio@ac.upc.edu
Ghent University, Gent, Belgium
piet.demeester@intec.ugent.be

Abstract. Aiming at creating a dynamic between elaboration, realization, and validation by means of iterative cycles of experimentation, Future Internet Research and Experimentation (FIRE) projects have been rapidly confronted to the lack of systematic experimental research methodology. Moreover, the “validation by experimentation” objective involves a broad spectrum of experimentation tools ranging from simulation to field trial prototypes together with their associated measurement tools. As experimental measurement results and corresponding tools play a fundamental role in experimental research, devising a systematic experimentation and measurement methodology becomes thus crucial for experimental research projects to achieve this objective for their various realizations, including protocols, systems and components. In turn, and in order to meet scientific validity criteria, the measurement results obtained when performing experimental research implies the availability of reliable and verifiable measurement tools, including on-line measurement data analysis and mining.

Keywords: measurement, experimental research, methodology, criteria

1 Introduction

The Future Internet Research and Experimentation (FIRE) initiative aims to realize a “research environment for investigating and experimentally validating highly innovative and revolutionary ideas” towards new paradigms for the Internet by bridging multi-disciplinary long-term research and experimentally-driven large-scale validation. FIRE foundational objectives were:

- Creation of a multi-disciplinary, long term research environment for investigating and experimentally validating highly innovative and revolutionary ideas for new networking architectures and service paradigms;

- Promotion of experimentally-driven yet long-term research, joining the two ends of academy-driven visionary research and industry-driven testing and experimentation, in a truly multi-disciplinary and innovative approach;
- Realization of a large scale European experimental facility, by gradually inter-connecting and federating existing and new “resource clusters” for emerging or future internet architectures and technologies.

These objectives further evolved toward the inception of experimentally-driven research as a visionary multidisciplinary investigation activity, defining the challenges for and taking advantage of experimental facilities. Such investigation activity would be realized by means of iterative cycles of research, oriented towards the design and large-scale experimentation of new and innovative paradigms for the Internet - modeled as a complex distributed system. The refinement of the research directions should be strongly influenced by the data and observations obtained from experiments performed at previous iterations thus, being “measurement-based” which in turn requires the specification of the relevant criteria and metrics as well as their corresponding measurement tools.

The rationale was thus clear: create a dynamic between elaboration, realization, and validation by means of iterative cycles of experimentation. The realization however was already less obvious and the validation objective rapidly confronted to the lack of systematic experimental research methodology and underlying measurement methodology applicable to computer communication and networked systems. Moreover, the “validation by experimentation” objective opens a broad spectrum of experimentation tools (ranging from simulation to prototype of real systems) and measurement tools. The selection of the experimentation tools depends itself on 1) the object of experimentation (corpus), 2) the nature and properties of the results, and 3) the cost function that itself depends on complexity, experimental and running conditions but also on the level of abstraction (referred to as “realism”).

Our main argument are that the systematic experimental validation of the targeted “elaboration and realization” can be described by a continuum that requires a broader set of tools: starting from more abstract tools (not only because their resulting cost is lesser but also because such tools produces results verifying all conditions explained in this Chapter) followed by the progressive addition of realism as part of the experimented system to ultimately reach the so-called field trials with real systems. The following sections detail the dependencies with respect to this experimentation chain and the set of criteria that experimental results shall satisfy in order to ensure the scientific validity of the results this chain produces.

Moreover, as measurement results and tools play a fundamental role in experimental research, devising a systematic measurement methodology becomes critical for the various experimental research projects to achieve the objective of validation by experimentation of their numerous realizations, including protocols, systems and components. In turn, the measurement-based experimental validation of

the targeted “elaboration and realization” involves a very broad set of measurement tools to obtain measurement results. In order to meet scientific validity criteria, these measurement results obtained when performing experimental research require the availability of reliable and verifiable measurement tools, including on-line measurement data analysis and mining.

2 Experimental Research Methodology

Computer communication networks are characterized by two fundamental dimensions: i) *distribution*: of a large number of dynamically interacting (non-atomic often complex) components, and ii) *variability*: the spatio-temporal variation of their inner properties that in turn influence these interactions. A couple of examples would better describe the landscape: autonomic networking is the transposition of the autonomic computing concept in the communication space, and network “virtualization” is the transposition of the abstraction concept of object-oriented programming in the networking space. More, the dynamic nature of these interactions results in modifying its scaling properties of the individual components besides modifying the properties of the global system. Many other examples can be cited, the fundamental observation is that no experimental model actually exists – or more precisely – the complexity of the resulting system makes its modeling a research discipline on its own.

However, this does not mean or imply that an experimental research methodology could not be defined based on i) a broader set of tools ranging from simulation to experimentation of prototypes/real systems and ii) experience gathered from practicing various and large experiments in the computer communication/networking disciplines. Such methodology would include the following steps (part of each iteration):

1. Specification of the functional and performance objectives, (technical and non-technical) constraints, and description of expected results
2. Definition of the relevant functional and performance criteria and metrics
3. Description of the *modus operandi* including configuration, initialization, running conditions, and (iterative) procedure(s) to be executed
4. The reporting on observations and the resulting analysis as well as the feedback after performing each iteration before reaching (partial) conclusion.

The following sections describe this methodology in more details.

2.1 Functional and Performance Evaluation

Functional and performance evaluation typically involves the specification of 1) Functional and performance model, 2) Functional and performance measurements and metrics, and 3) Functional and performance results analysis. We describe in the

following the experimentation methodology for performance evaluation and refer to [1] for what concerns for the functional evaluation methodology.

- **Performance model:** the specification of a performance model defines the significant aspects of the way by which a proposed or actual system operates in terms of resources consumed, accessed, scheduled, etc. together with the various delays resulting from processing and/or physical/hardware limitations (such as bandwidth, access latency, etc.). A performance model provides useful information on how the proposed model vs. actual system will or does actually work. Based on the information contained in the performance model, the interpretation of the execution of the model (by means of simulation or emulation) provide further insight into the system's behavior, and can be used to identify where the model design is inadequate.
- **Performance measurement:** many performance metrics may be used for this purpose, including computational complexity, memory-space complexity, time complexity (convergence), communication complexity, etc. but also combination of metrics including adaptation cost (which combines communication and computational complexity).
- **Performance analysis:** includes i) the examination of the performance measurement results obtained for the proposed model against the criteria associated to each performance metric, ii) the comparison of the performance measurement results obtained for the proposed model against those obtained for the reference model, and iii) the conformance of the performance measurement results obtained for the proposed model against the results as determined and/or estimated by the associated performance model.

2.2 Experimental Evaluation Methodology

Experimental performance evaluation requires to specify a theoretical model (of the system under study) from which a performance analysis can be performed. Using the feedback from this performance analysis a behavioral/conceptual model is then built that enables the development of an experimental model being either a simulation or an emulation model. A simulation model aims at characterizing the working of the modeled process or system over time by examining a range of behaviors that are similar or analogous to a real world system. On the other hand, an emulation model imitates the externally observable behavior to match an existing real system functionally close enough so that it can be substituted to the real system while internal state of the emulation mechanism does not have to accurately reflect the internal state of the real system which it is emulating. Simulation and emulation are further detailed in Section 4.

The experimental model can then be converted into a computational model. Measuring the metrics on the execution of this model provides the needed information to compare the obtained results with those of the theoretical model. Fig.1 depicts the flow chart used for the purpose of systematic performance evaluation (and analysis).

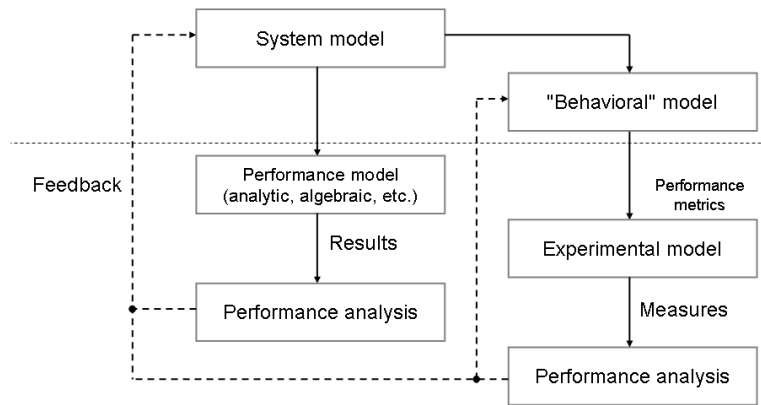


Fig. 1. Performance evaluation and analysis methodology

The (typically iterative) process to build and develop such experimental model is the following:

- Determine the goals and objectives of the experiment
- Build a conceptual model including state variables, which variables are static and dynamic, for the latter are variations continuous and discrete, etc.
- Convert the conceptual model into a specification model of the experiment being a simulation or emulation experiment; the specification typically describes the experimented procedures (pseudo-code) and data structures; the experimented procedures can be either simulation and/or emulation procedures or even prototype(s).
- Convert the specification model into a computational model, i.e., executable computer program(s). The selection of the programming language is part of this step and consists in determining whether a general-purpose programming language or a special-purpose language would be used to develop this/these program(s). Note that in case of simulation, the term program is frequently used to refer to the computational model itself.
- Verify (verification process): Ask the question: Did we build the model right?
 - Determine whether the computational model executes as intended by the specification model.
 - Determine whether the computational model implements the assumptions made about the behavior of the real systems (as transposed in the behavioral model)
 - Verification techniques include: tracing/walk-through, continuity tests (sensitivity tests, i.e., slight change in input should yield slight change in output, otherwise error), degeneracy tests (perform execution with extremes values,

e.g., lowest and highest), and consistency tests (similar inputs produce similar outputs).

- Validate (validation process): Ask the question: Did we build the right model?
 - Determine whether the conceptual model is representative of the actual system being analyzed or not. Can the conceptual model be substituted, at least approximately for the real system?
 - The validation process also involves determining whether the computational model is consistent with the actual system being analyzed.

Once the executable computer program(s) are verified and validated, experimental studies can be performed. Performance evaluation studies include the following steps:

- Design the experiments:
 - Determine the input parameters that should be varied and their interval as well as the initial conditions for proper initialization (note that some characteristics of the environment may need to be included in the experiment if not accounted as part of the simulation model); this step is critical as it provides means for decreasing the run time of the simulation but still may not provide confidence for stable conditions.
 - Determine the variables to measure (at which frequency, upon which class of event (event-driven), etc. taking into account the tradeoff between too much data (that would in turn require the use of techniques for reducing the amount of collected to a usable form) and too little data (that would in turn introduce the need for representing data by statistical distributions);
 - Determine the execution time taking into account the tradeoff between the resource consumption (by very long runs) and amortization of effects of resulting from transient state will be amortized.
- Execute the program(s) and record tuples of the form {<initial_condition; input>; <running_condition; output>} referred to as observations or data.
- Analyze the output of the program execution towards production of results. Such analysis comprises the following:
 - Results verification (correctness): test whether the results obtained are in accordance to the assumptions made about the behavior of the real systems (as transposed in the behavioral model);
 - Input validation (representativity): validate assumptions about input parameter values and distributions. This step is often associated to the output validation phase;
 - Results/output validation (representativity): test whether the results obtained are representative of those obtained either by real systems or theoretical model of the system.

3 Experimental Results Criteria

Let's now proceed with the definitions of the criteria that experimental results shall satisfy when performing experimental evaluation in order to ensure that scientific

validity of the results these experiments produce. If we model the corpus of a given experiment (or experimental corpus¹) by a function F , with input vector of variables x_1, \dots, x_n and parameters e_1, \dots, e_m such that its output $F(x_1, \dots, x_n | e_1, \dots, e_m) = y$, then the following properties shall hold:

3.1 Reliability

Reliability is defined as the probability that system or component will perform its intended function during a specified period of time under stated conditions. Referring to Fig.2, it means that during the pre-defined time interval $[t_1, t_n]$ the output of the experimental corpus model $F(x_1, \dots, x_n | e_1, \dots, e_m)[t_k] = y[t_k]$ exists $\forall k \in [1, n]$ within a pre-defined range $[y_1, y_p]$ of valid output.

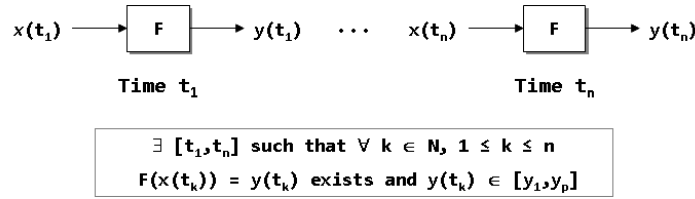


Fig. 2. Experimental criteria: reliability

3.2 Repeatability

Referring to Fig.3, repeatability means that $\forall k \in N_0$, if $(x_1, \dots, x_n | e_1, \dots, e_m)[t_{k-1}] = (x_1, \dots, x_n | e_1, \dots, e_m)[t_k]$ together with the condition that $F(x_1, \dots, x_n | e_1, \dots, e_m)[t_{k-1}] = y[t_{k-1}]$ and $F(x_1, \dots, x_n | e_1, \dots, e_m)[t_k] = y[t_k]$ exist, then $y[t_k] = y[t_{k-1}]$. The term reliability is thus characterized by persistence of the output in time.

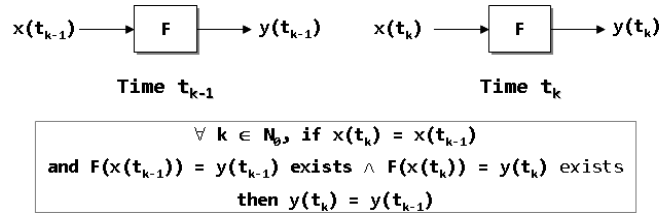


Fig. 3. Experimental criteria: repeatability

¹ The term experimental corpus refers to the main object of an experiment.

3.3 Reproducibility

Reproducibility means that the experimental model $F(x_1, \dots, x_n | e_1, \dots, e_m)$ can be executed at the same time (simultaneously) on different experimental systems $u, v \in S$ (\equiv experimental system set) and produces the same output if the model input at both systems is identical. Reproducibility is thus characterized by persistence of the output in space. We refer to Fig.4 for a formal definition of reproducibility.

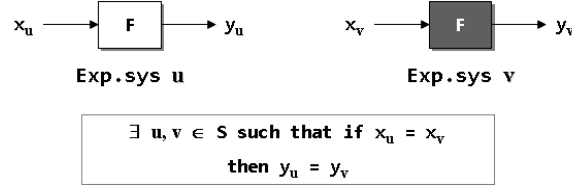


Fig. 4. Experimental criteria: reproducibility

3.4 Verifiability

Referring to Fig.5, verifiability means that we can find independently a formal model $H: x(t) \rightarrow H(x(t))$ corresponding to the experimental model $F: x(t) \rightarrow y(t) = F(x(t))$ such that at time t_k , the output of the function F can be confirmed against the output of the formal model H , i.e., referring to Fig.5, $H(x_1, \dots, x_n | e_1, \dots, e_m)[t_k] \in [y(t_k) - \varepsilon; y(t_k) + \varepsilon]$ with $\varepsilon \ll 0$, where $y(t_k) = F(x_1, \dots, x_n | e_1, \dots, e_m)[t_k]$.

Note that in practice, one aims at finding a formal model H such that the output of the function F complies with the output of the model H at any time $t_k, \forall k \in [1, n]$, time interval defining the duration of the experiment.

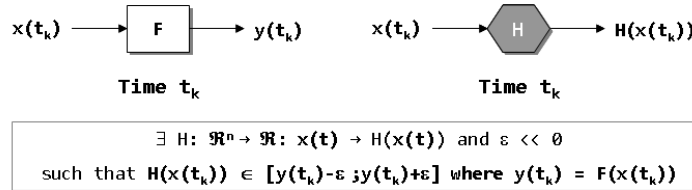


Fig. 5. Experimental criteria: verifiability

In order to ensure verifiability, reliability, repeatability, and reproducibility of the experimental results produced, one shall characterize the output of experimentation. Meeting these criteria implies in turn to control the parametrization, the input and output, as well as the running conditions of the conducted experiments. On one hand, verifying the repeatability, reproducibility, and reliability criteria enables

generalization of the experimental results produced. On the other hand, ensuring verifiability of these results increases their credibility (results can be "explained").

In the present context, it is also important to underline the fundamental distinction between verification (verifiability) and validation (validity). Verification means that the experimental model output should satisfy the formal model output (e.g. computational model). On the other hand, validity is formally defined as follows: a proposition A is valid if $H(A) = \text{TRUE}$ for any model H of A; thus, in experimental research, we can only verify satisfiability (proposition A has at least one model H for which $H(A) = \text{TRUE}$). Note that in propositional logic, one usually verifies validity by applying the following theorem: a proposition A is valid if and only if its opposite (negation) can not be satisfied. Henceforth, the best we can hope concerning "verification" is to find at least one model H of A that verifies the same "output" as the realization F of A by experiment: $H(A) = F(A)$. If this is the case, then one does indeed satisfy the initial proposition (it is verified by one model) but not validate it (the proposition is not verified for any model). Also, one constructs (independently) a model to verify that the output of the experiment F satisfies to the output of the model H. Thus, one does not verify the conformance of experimental execution against the specification of the experimented system but the consistency of the experimental output against a computational model of the experiment drawn independently from it. Verifiability is thus not synonym of conformity test or conformance test against the specification of the underlying experimented system. Finally, it is interesting to observe from its definition that verification is the formal complement to experimentation (instead of positioning experimentation as the complement to the theoretical model).

Note: scalability is often cited as a criterion to be met by the experimental corpus. The scale of a system is measured by the rate \times state \times size that the system can sustain when running using a given number of resource units (for processing and storage). Networking systems can thus only scale indefinitely if and only if the rate of change of the state set, the number of states and the size of each state are independent of the global properties of the environment into which the system is operating. It is thus fundamental to mention that the scale of an experimental facility (the number of resource units and their distribution) does not determine the scalability properties of the corpus. However, the scalability properties of the experimented corpus determine the number of resource units that are locally required to be executed at a certain scale. Thus, such experiment can be performed to i) verify a pre-estimated level of scaling of the experimented corpus and/or ii) iteratively determine the scale property of the corpus with the risk that the dependency on the global properties could never be found (hidden dependencies, correlations, non-linearities, etc.). Hence, only the former leads to verifiable experiments. In other words, a large-scale facility can only help verifying scaling properties but not determine these properties. Further positioning the role of so-called large-scale testbeds/experimental facilities is outside the scope of this Chapter.

4 Experimental Tools

Different experimental tools can be used. Their selection is neither arbitrary nor religious: it depends on the experimental objective and maturity of the experimental corpus. Nevertheless, each of them needs to ensure that the experimental criteria defined Section 3 are satisfied. However, it is clear that fulfilling these criteria does not come at the same cost for the same level of abstraction. We can distinguish three types of abstraction: i) abstraction of the network/shared infrastructure (network resource consumption model, processing model, etc.), ii) abstraction of the system (processing/memory resource consumption model, computation model, etc.), and iii) abstraction of the environment (traffic model, application model, user/behavior model, etc.). To each (non-atomic) element of this partition of the abstraction space, we can associate a level of realism when the abstraction is replaced by a “real” entity. Without entering into the debate of reality or what reality actually represents or means, we simply consider here a real system as an instantiation of the experimented component models at the hardware and/ or software substrate level depending on the expected level of performance.

4.1 Simulation

Following Shannon [2], simulation is the process of designing a model of a real system and conducting experiments with this model for the purpose either of understanding the behavior of the system or of evaluating various strategies (within the limits imposed by a criterion or set of criteria) for the operation of a system. The execution of the model X is said to simulate another system Y when the internal working processes of X can be described by a mathematical and/or procedural model known to best represent the actual working processes of Y . Ingalls [3], further defines simulation as the process of designing a dynamic model of an actual dynamic system for the purpose either of understanding the behavior of the system or of evaluating various strategies (within the limits imposed by a criterion or set of criteria) for the operation of a system.

Simulation can thus be seen as the process of exercising a model to characterize the working of the modeled entity process, or system over time by examining a range of behaviors that are similar or analogous to a real world system. Simulations are never, by definition, complete. Simulation is one of the most widely used techniques for i) understanding, characterizing and analyzing the behavior of complex systems, ii) construct theories or hypotheses that account for the observed behavior, iii) use the model to predict future behavior, that is, the effects that will be produced by changes in the system. Simulations may be deterministic or stochastic, static or dynamic, continuous or discrete.

4.2 Emulation

Emulation is the process of imitating the outwardly observable behavior to match an existing real system functionally close enough so that it can be substituted to the real system. The internal state of the emulation mechanism does not have to accurately reflect the internal state of the real system which it is emulating. A system X is said to emulate another system Y if the behavior of X is exactly the same as that of Y (same output for same input under similar conditions) but the mechanism to arrive at the output (from the input) is different. Emulation is generally used when we don't exactly know the internal mechanism of the original system but are familiar with the input/output pattern. Performing emulation enables thus to imitate the function of the real system, as by modifications to hardware or software that allow the imitating system to accept the same data, execute the same programs, and achieve the same results as the imitated system without exactly reflecting the inner functioning of the real system. The emulation is "complete" if all the interfaces are present, and the resulting observed behavior matches that of the real world device.

An emulator can thus be defined as a model of a system which will accept any valid input that that the emulated system would accept, and produce the same output or result. Hence, across a well defined interface an emulator is indistinguishable from the real world equivalent (except in performance).

4.3 Simulation and Emulation are Complementary

In general, spatial measures are more easily achieved in simulation environments whereas temporal measures are more easily achieved by means of emulation. Simulations (in particular, simulation by discrete event) is well suited for experiments involving spatial metrics, structure, and dimensions as it enables handling of large-scale topology cases and produces results that are easier to tune, reproduce, and compare. On the other hand, emulation has the advantage of being more representative of the actual execution of procedures thus providing valuable insight in temporal metrics and behavior (although on smaller scale and at a higher cost). Moreover, emulation experiments enable to check the realism of simulation results and simulation experiments to extend the applicability of emulation results. Indeed, results obtained by means of simulation and emulation experiments are complementary when the experimental scenarios are commonly specified and their execution adapted (without introducing any bias) to the simulation or emulation environment.

Emulation experiments can lead to reproducible and repeatable results but only if their "conditions" and their "executions" can be controlled. Realism can thus be improved compared to simulation (in particular for time-controlled executions of protocol components on real operation system). Nevertheless, such experiments are more complex and time consuming to configure and execute; performance evaluation is possible if the experimental platform comprises a "sufficient number" of machines

(representative of the experiment to conduct). Emulation still requires synthetic network conditions (models) if executed in controlled environment and either injecting real traffic or replay traffic traces (not that even when available "spatial distribution" of traffic is available remains problematic to emulate because the spatial aggregation of address prefixes is not necessarily as the routing tables are often not provided together with traffic traces).

Stepping into real system experimentation increases cost but increases realism. As such the loss of control of experimental conditions in such systems raises the issue of persistence of the properties observed earlier in the experimentation chain. In particular, these properties shall already be determined by the earlier experimental stages (leaving them intrinsically part of experimental research activities).

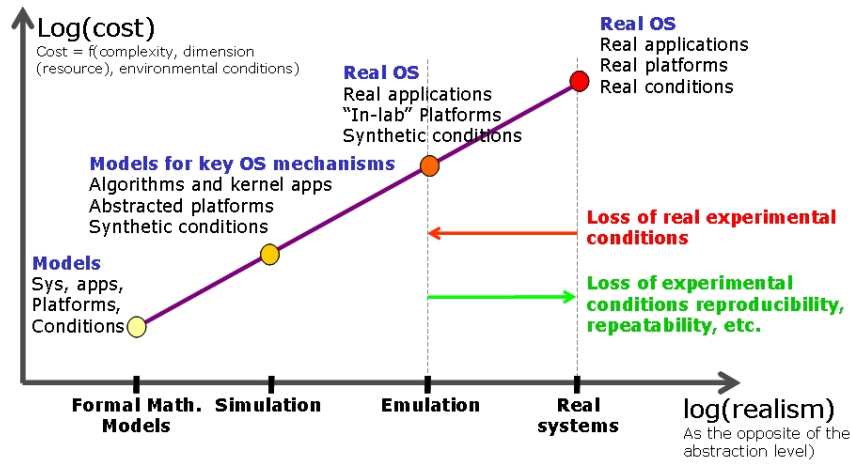


Fig. 6. Experimental criteria: verifiability

In this context, the validation of a new routing algorithm for instance, would be better conducted on a simulation platform (after formal verification) not only because their resulting cost is lesser but because such tools produces results verifying all conditions explained here above. Afterwards, progressive addition of realism as part of the experimented system would consist in instantiating the execution stratum (remove the system abstraction) in order to perform emulation experiments.

5 Measurement Methodology

Measurement refers to metrology which is defined by the International Bureau of Weights and Measures (BIPM) as "the science of measurement, embracing both experimental and theoretical determinations at any level of uncertainty in any field of science and technology." The term metrology includes all aspects of measurement

(theoretical and practical) [4] [5]: starting from the "principles of measurement", which represent the scientific basis for measurement, the "method of measurement" (logical sequence of operations) is instantiated by a measurement (set of operations).

5.1 Measurement Objectives

As documented in [6], measurements aim at determining not only i) the value of a quantity but also at determining, ii) the distributions of quantities in time, in space, and in time and space, iii) the mathematical representations of quantities or their distributions, iv) the relations between quantities, their distributions or representations, and v) the parameters of such relations. The results of measurements of types (i) and (v) are expressed in terms of numbers. The results of measurements of types (ii), (iii) and (iv) may have the form of numbers, series of numbers, functions or series of functions -given in tables-, or analytically. When measurements of type (v) are considered, then the parameters of relations between quantities are often treated as new quantities (e.g., resistance, inductance, capacitance), but the diversity of the investigated relations (e.g., non-linearity, dependence on frequency) breaks the quantitative concept of measurement.

From this perspective, measurement-based experimental research aims at complementing the rigorous performance analysis and simulation-based evaluation. The results are more realistic and can contribute to validate and to fine tune the execution of algorithms. A large variety of realistic topologies, mobility profiles, and traffic patterns is required. Novel network parameters as well as performance monitoring measures (and their trade-offs) arise. Ad hoc approaches are useful but there is a need to converge to widely accepted, common integrated measurement methods, systems and tools.

5.2 Measurement Process

The measurement process is instantiated in a measurement procedure having the *measurand* (quantity that is to be measured) as its inputs, the control variables, and the output representing the *measurement results*. The measurement process comprises 3 distinct steps: 1) design of a measurement method, 2) application of measurement method rules, and 3) analysis of the measurement results. To carry out a measurement task, an experimenter should design and execute a measurement procedure (corresponding to the measurement function μ) which consists of a set of operations, specifically described, for the performance of a particular measurement according to a given measurement method. Note that the results of the measurement can be influenced by external quantity during the measurement process. As experimental testbeds are of different nature (wired, wireless, different hardware, etc.) and offers different level of control of the experimental initial and running conditions the measurement method will have to be specified in order to ensure these external quantities can be identified.

As experimental research in the context of FIRE is by nature distributed, the measurement process shall account for the distributed nature of the experimental environment and the distributed nature of the experimented corpus. This poses additional challenges that should be met by the measurement tools used in the measurement process, including control of the properties of the measurement tools and their calibration, measurement timing and synchronization (timestamps), as well as the measurement granularity, sampling and representativeness.

5.3 Measurement Results Analysis

In order to ensure that measurement results can be systematically analyzed, the control of the experimental execution conditions together with the following elements have to be considered:

- Specify performance analysis methodology together with the necessary mathematical tools to be able to perform data analysis and mining tasks on experimental data coming from various monitoring points (from single or multiple testbeds). This objective also covers specification the necessary mathematical tools to analyze the sensitivity of the performance measures to changes in the *experimental model* parameters. Sensitivity analysis attempts to identify how responsive the results of an experimental model are to changes in its parameters: it is an important tool for achieving confidence in experimentation and making its results credible. Sensitivity analysis quantifies the dependence of system behavior on the parameters that affect the modeled process and in particular its dynamics. It is used to determine how sensitive a model is to changes in the numerical value of the model parameters and changes in the model structure.
- Specify distributed performance monitoring system (while) allowing experimenters to choose the best tool(s) for their experimentation.
- Define a standard experiment description together with a control interface and wrap the measurement tools within this interface. This standard interface will focus on providing a common programming interface to describe every aspect of a networking experiment but will also attempt to provide robust experiment monitoring and management facilities and will integrate with the data analysis and data mining tools developed as described here above (cf. first bullet point). Note that sensitivity analysis of the reliability, the performance, and the performability of the monitoring system is a complementary objective.

6 Measurement Criteria

Measurement results obtained by means of experiments have to verify certain properties and criteria. These properties and criteria mainly include reliability, repeatability, reproducibility, and verifiability. These criteria are indeed the same as those that experimental results shall satisfy (see Section 3) but in the present case they are applied in order to characterize the scientific validity of the measurement results

these experiments aim at producing. In turn, they constraint the experimental corpus and research methodology but also determine the fundamental properties and criteria that shall be met by the measurement tools used to perform the measurements implied by these experiments.

6.1 Reliability

Reliability is defined as the probability that the measurement function μ performs its intended measures (output) during a specified period of time under stated conditions. More formally, referring to Fig.7, where the experimental corpus is modeled by a function F with input vector x and output $y = F(x)$, reliability is verified when $\exists [t_1, t_n]$ and $\varepsilon \ll 0$ such that $\forall k \in \mathbb{N}, 1 \leq k \leq n, \mu(y(t_k)) = \mu(F(x(t_k))) \wedge y(t_k) \in [\mu(y(t_k)) - \varepsilon, \mu(y(t_k)) + \varepsilon]$, where $y(t_k) = F(x(t_k))$.

Reliability implies as a minimum requirement that the components of the experimental corpus remain operational (i.e., do not fail or halt) during this time period. Furthermore, measurement results are reliable if they remain consistent (within a certain well-defined range) during that period.

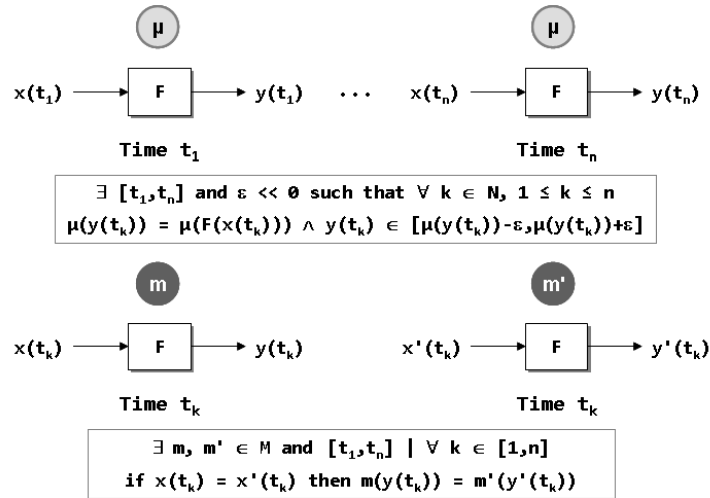


Fig. 7. Measurement criteria: reliability

In practice, in order to assert the reliability of a given measurement tool implementation m ($\in M \equiv$ measurement program set) of a given measurement function μ , it is common to compare the results of measurements produced by m with those obtained for the same time period by means of another implementation m' ($\in M$) of the same function μ . Referring to Fig.7, $\exists m, m' \in M$ and $[t_1, t_n]$ such that $\forall k \in [1, n]$, if $x(t_k) = x'(t_k)$ then $m(y(t_k)) = m'(y'(t_k))$.

6.2 Repeatability

Repeatability is a temporal criterion associated to measurement results. This term is used when multiple execution of a given experiment (repetition) using the same configuration, running conditions, and input yields the same output. Correct experimental method and usage of models, execution of algorithms, and output processing are required in order to guarantee the repeatability of measurement results.

More formally, referring to Fig.8, repeatability is verified when the following condition is met $\forall k \in \mathbb{N}, k \geq 1$, if $x(t_k) = x(t_{k-1})$ then $\mu(y(t_k)) = \mu(y(t_{k-1}))$, where $y(t_k) = F(x(t_k))$ and $y(t_{k-1}) = F(x(t_{k-1}))$.

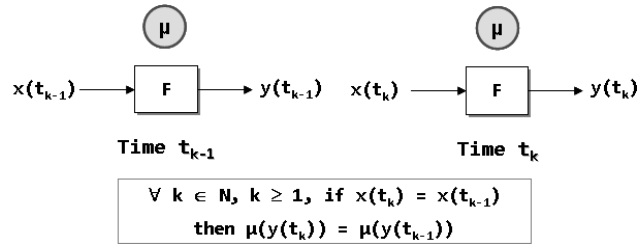


Fig. 8. Measurement criteria: repeatability

6.3 Reproducibility

Reproducibility is a spatial criterion associated to measurement results that can be obtained when a given experiment performed on a given experimental system u ($\in S \equiv$ experimental system set) is replicated over a similar but different experimental system v ($\in S$). This can mean different experimental platform, operating system, etc.

Typically, reproducibility comes into play when a third party performs the same experiment to determine the scientific validity of the output of an experiment. More formally, referring to Fig.9, reproducibility is achieved when $\exists u, v \in S$ such that if the input vector $x_u = x_v$ then $\mu(y_u) = \mu(y_v)$, where $y_u = F(x_u)$ and $y_v = F(x_v)$.

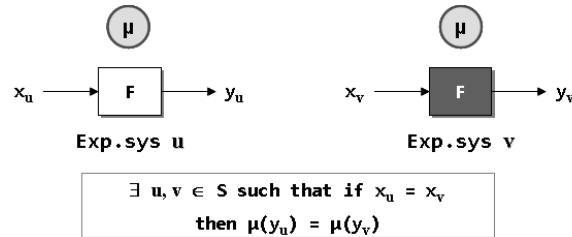


Fig. 9. Measurement criteria: reproducibility

6.4 Verifiability

The results of experimental measurements are verifiable if the output of the experimental corpus modeled by the function $F: x(t) \rightarrow y(t) = F(x(t))$ can be confirmed against a formal model $H: x(t) \rightarrow H(x(t))$; implying that the measurement results shall comply with the output of the model H (output described as a function of the input vector and the experimental parameters).

Referring to Fig.10, measurement results are verifiable if there exists a formal model $H: \mathcal{R}^n \rightarrow \mathcal{R}: x(t) \rightarrow H(x(t))$ and $\varepsilon \ll 0$ such that at time t_k , $H(x(t_k)) \in [\mu(y(t_k)) - \varepsilon; \mu(y(t_k)) + \varepsilon]$, where $\mu(y(t_k)) = \mu(F(x(t_k)))$. One often considers that verifiability is achieved by comparing the results of an experimental measurement against a reference system RS (assumed as representative of the real system): $\exists u \in S$ and $\varepsilon \ll 0$ such that if $x_u = x_{rs}$ then $y_{rs} \in [\mu(y_u) - \varepsilon; \mu(y_u) + \varepsilon]$.

Achieving verifiability for a representative sample (to avoid sampling bias) of unbiased measurement results whose size is determined so as to reduce the sampling error (and satisfy a given confidence interval and level given the finite but often large number of available results) enables in turn to generalize the conclusion(s) that can be drawn from experimental measurements.

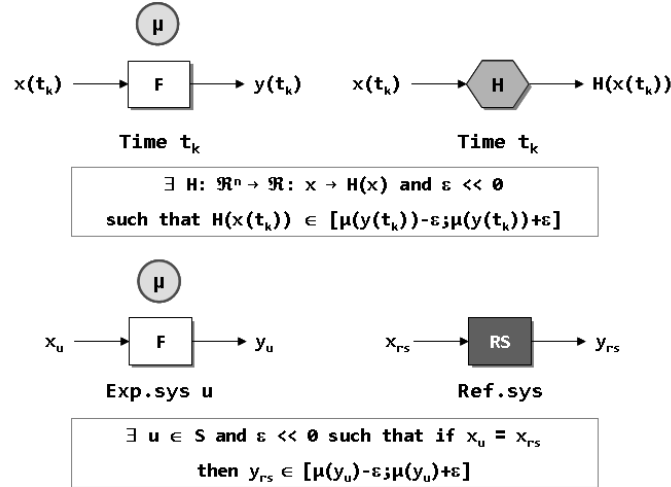


Fig. 10. Measurement criteria: verifiability

7 Conclusion

Starting from the initial objectives of the FIRE initiative and its associated objective of dynamic between elaboration, realization, and validation by means of iterative cycles of experimentation, this Chapter positions measurement-based

experimental research as a continuum: starting from more abstract tools (not only because their resulting cost is lesser but because such tools produces results verifying all criteria explained in Section 3) followed by progressive addition of realism as part of the experimented system to ultimately reach the so-called field trials with real systems. The addition of realism at increasing cost (resulting from the increasing complexity) is the main purpose of performing experimentation by means of emulation or real systems. However, achieving verifiable, reliable, repeatable, and reproducible emulation results at best cost-complexity (thus experimentation time) can not be achieved if experiments are limited to random trials on emulated platforms. Indeed, emulation experiments can lead to reproducible and repeatable results but only if their "conditions" and their "executions" can be strictly controlled. Realism can thus be improved compared to simulation (in particular for time-controlled executions of protocol components on real operation system). Nevertheless, such experiments are more complex and time consuming to configure and execute.

In order for measurement-based experimental research to reach this objective: 1. This paper has proposed a systematic experimental research methodology which needs to be commonly shared and applied by projects in order to ensure the scientific validity of the experimental results they produce (otherwise, leaving them as only purpose the proof of executability of implementation instances of experimental corpus); 2. The validation by experimentation of the targeted “elaboration and realization” being actually a continuum, it requires in turn a set of well-defined experimentation tools to systematically implement this methodology starting from more abstract tools such as simulation (not only because their resulting cost is lesser but because such tools produces results verifying all conditions explained in this Chapter) followed by progressive addition of realism as part of the experimented system, e.g., by means of emulation, to ultimately reach the so-called field trials with real systems; and 3. As experimental measurements play a fundamental role in experimental research, developing a systematic measurement methodology is crucial to achieve the objective of validation by experimentation of the various project outcomes including protocols, systems and components. Moreover, in order to meet scientific validity criteria, the measurement results obtained when performing experimental research implies the availability of reliable and verifiable measurement tools providing repeatable and reproducible measurement results. Finally, as the amount of collected data increases (due to the scale of experiments), the development of measurement tools shall not be limited to actual measures and their collection but also on-line analysis and mining of measurement data.

Acknowledgements

This research work is conducted by the EULER Project (Grant No.258307) part of the Future Internet Research and Experimentation (FIRE) objective of the Seventh Framework Programme (FP7) funded by the European Commission (EC).

References

1. EULER FP7 Project, *Performance objectives, evaluation criteria and metrics*, Technical report. Available at: <https://www-sop.inria.fr/mascotte/EULER/wiki/pmwiki.php/Main/Deliverables>
2. Shannon, R.E., *Systems Simulation the Art and Science*, Prentice-Hall, Englewood Cliffs (NJ), USA, 1975
3. Ingalls, R.G., *Introduction to Simulation*, Proceedings of the 34th Winter Simulation Conference, San Diego (CA), USA, December 2002.
4. International Organization for Standardization (ISO), *International vocabulary of metrology - Basic and general concepts and associated terms (VIM)*, ISO/IEC Guide 99:2007.
5. Abran, A., and Sellami, A., *Initial Modeling of the Measurement Concepts in the ISO Vocabulary of Terms in Metrology, Software Measurement and Estimation*, Proceedings of 12th International Workshop on Software Measurement (IWSM), Magdeburg, Germany, Oct. 7-9 2002, Shaker-Verlag, Aachen, 2002.
6. Fiok, A.J., Jaworski, J.M., Morawski, R.Z., Oledzki, J.S. and Urban, A.C., *Theory of measurement in teaching metrology in engineering faculties*, Warsaw University of Technology, PL-00-661 Warsaw, Poland.